学校编码：10384　　分类号_____密级_____
学号：X2013230260　　UDC_____

厦门大学

工 程 硕 士 学 位 论 文

基于 Teradata 的银行数据仓库的模型研究与设计

Research and Design of Data Warehouse Model in Bank Based on Teradata

杜 骞

指 导 教 师：王 备 战 教 授
专 业 名 称：软 件 工 程
论文提交日期：2015 年 4 月
论文答辩日期：2015 年 5 月
学位授予日期：　　年　　月

指 导 教 师：_____
答辩委员会主席：_____

2015 年　月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘要

企业数据仓库的建立对于企业管理具有重要的意义。不同企业须根据其自身特点和现实环境来选择建设数据仓库的路线和方法。

企业级数据仓库发展至今，已经进入一个全面发展阶段。以银行业的企业级数据仓库为例，一个成熟的企业级数据仓库需要具备至少海量数据存储能力(PB级)、数据模型能力(数据按业务主题进行归类如：客户、帐户、会计、交易明细)、大数据量分区能力、大批量处理的计算能力、数据质量的控制能力、高并发处理能力、报表数据存储能力和非结构化数据管理能力等特点。

本文着眼于数据仓库数据模型的研究，着重介绍模型设计客户化的方法，结合本人工作参与的某银行数据仓库建设项目中使用的模型设计方法和经验，为未来类似项目中数据仓库的构建奠定了良好的技术基础。

本文从数据仓库的研究背景入手，介绍了目前国内银行数据仓库建设的现状，以本人实际参与的项目为依托，展开基于 Teradata FS-LDM 的数据仓库模型的研究。首先介绍了数据仓库的相关概念，通过对多维模型和关系模型两种建模方法的比较引出数据集市的相关概念，而后介绍了 Teradata 数据仓库作为当今使用最为广泛的数据仓库的优势，以及依据其成熟的数据仓库可扩展方法论构建数据仓库的基本框架所包含的内容。最后以理论结合实际的方式阐述了数据仓库模型的设计方法，重点介绍了模型设计客户化的工作流程、工作内容与工作经验。最后对本论文进行了总结与展望，并对数据仓库未来的发展趋势进行了展望。

关键字：Teradata；银行数据仓库；数据模型

Abstract

The establishment of data warehouse plays an important role in the corporate management. Different enterprises should choose different ways and approaches to establish their data warehouse.

The enterprise-level data warehouse has come to an all-round development stage after a long-term development. Taking the data warehouse of the banking industry as an example, a mature enterprise-level data warehouse should at least have the following features: Massive data storage (Calculated by PB), Data model (Classify data according to different businesses, such as client, account, accounting, transaction details etc.), Large data volume partition, Mass-processing computing power, Control over data quality, High concurrent processing, Report data storage and Unstructured data management.

This dissertation focuses on the studies of data warehouse and data mode. Particular attentions are paid to the method of designing customized model. The personal experience of the data warehouse designing method and experience in a bank lays good technological foundation to the similar data warehouse establishment in the future.

The dissertation introduces the present situation of the data warehouse in domestic banks beginning with the research background and makes the research based on the Teradata FS-LDM model depending on the personal-participated project. Relative definitions of the data warehouse are explained at first, and the concept of data mart is introduced by comparing the two modeling approaches-multidimensional model and relation model. Then the advantages of Teradata data warehouse as the most widely used data warehouse are presented as well as the covered aspects of establishing data warehouse according to the extending method of the mature data warehouse. The designing method of the data warehouse model, especially the workflow, contents and experience of customized model designing, is illustrated by integrating theory with practice. In the end, conclusion of the dissertation is drawn and the future data

warehouse development is anticipated.

Key Words:. Teradata; Bank Data Warehouse; Data Model

目 录

Contents

第一章 绪 论

1.1 研究背景

随着企业信息化的不断深入，在日益激烈的市场竞争中，信息化建设的程度对企业的生存和发展起到了决定性的作用。伴随着企业的飞速发展与壮大，由业务发展而产生的数据也会不断膨胀。有研究表明，企业关注的数据比例不超过总数据的 5%。那么在这些庞杂的数据中，如何挖掘出蕴含商业价值、能够使经营决策者掌握充足的企业信息和降低决策风险的数据，成为了一个迫切需要解决的问题。而企业数据仓库的建立能够集中分散在企业各处的数据并按需整合，提供一致性较高的数据，为企业决策提供强大的数据支撑，从而很好的解决了这一问题。

数据仓库的概念始于上世纪 80 年代中期。早期，基于各业务部门的不同业务需求，企业中产生了不同的应用系统，在这些应用系统的产生的数据基础之上设计和构建出了核心业务的数据库系统，如财务管理、客户管理、销售管理等系统[1]。但是随着企业的发展，业务部门的扩充，企业对有效数据信息的获取需求也随之加大，分散的业务数据库和异构的数据源已不能满足企业对信息获取的需求，为满足企业对全局数据信息的获取需求，支持远景决策分析，需要一个可以覆盖所有业务部门的大型应用系统，即 ERP 系统，而基于 ERP 之上的数据仓库也随之诞生。与传统数据库不同的是，数据仓库是专门面向分析应用，面向企业高端决策层，将收集到的企业数据以维度集成而建立的一个企业数据中心，这些数据用来生成报表和用于查询，为企业提供战略决策上的数据支撑。

在银行领域中，通过数据仓库建立专业有效的信息管理体系尤为重要。近年来国内各银行都在致力于完善自身的信息化建设程度，规模较大较为完善的业务系统也在持续不断的开发中。但与外资银行相比，我们在数据分析和数据挖掘方向上还有比较大的差距，这个差距驱使国内银行提高了对自身信息技术的建设要求，为达到有效传递数据信息，实现数据共享和更广泛的应用，各银行纷纷建立了数据仓库。

1.2 研究现状

银行领域的数据仓库发展至今，已经进入一个全面发展阶段。一个成熟的银行数据仓库需要至少具备以下几个特点：

1. 海量数据存储能力(PB 级)
2. 数据模型能力(数据按业务主题进行归类如：客户、帐户、会计、交易明细）
3. 大数据量分区能力
4. 大批量处理的计算能力
5. 数据质量的控制能力（数据传输技术检核能力、数据横向检核能力）
6. 高并发处理能力（批处理计算与在线访问）
7. 报表数据存储能力
8. 非结构化数据管理能力（不同格式非结构化数据的导入能力，支持数据分级存储的能力，数据版本自动更新能力，非结构化数据调阅、下载和搜索挖掘能力，数据随应用需求集中和分散部署能力，非结构化数据全生命周期管理能力）

在银行数据仓库的构建过程中，模型建设可谓是重中之重。一般来说，模型分为逻辑模型和物理模型两部分。如果数据仓库逻辑模型设计的好，与业务领域结合的紧密，掌握好物理模型和逻辑模型的平衡点，银行对数据的有效利用率也会显著提升。随着国内各大银行对自身数据仓库构建需求的提升，业界已存在的成熟的逻辑模型，如 Teradata 的 FS-LDM（Financial Services Logical Data Model）为数据仓库的构建提供很好的蓝图和指引。Teradata 在数据仓库实施和金融领域方面经验丰富，基于 Teradata 的数据仓库 FS-LDM 是预先构建的逻辑模型，它是按照三范式原则设计的，因此利用它可以直接开始数据仓库的模型设计并可以运行在任何的数据平台上，它遵循了中性与共享、一致性、灵活性、粒度性的设计原则，遵循了面向主题的设计方法，为建立一个强有力的数据仓库奠定了重要的基础。同时它的灵活易扩展性，使银行在不断增加业务功能时不需要重钩整个数据仓库。

1.3 论文的来源和研究内容

对于银行而言，随着业务的发展和各个业务子系统的建立，信息归集和数据集成问题凸显，如对复杂的业务数据服务响应速度慢；依靠分散的报表系统获取的基础数据对分析决策的支持能力较低；缺乏统一的数据平台进行数据集成整合，难以为产品服务层和管理分析层提供一致性的数据服务。为了应对以上问题，银行须确立以数据仓库为基础的决策支持系统，建设数据集成平台，采用面向主题的仓库模型设计方法实现系统报表定制、即席数据查询等功能，重点实现了业务层面上的多维度分析，为业务部门提供所需的数据支持。

建立逻辑模型是数据仓库建模的第一步，也是为后续应用提供数据分析的基础。在逻辑模型设计方面，模型的客户化是设计的重点，在设计阶段根据具体的业务需求定义解决方案，根据具体的业务应用进行详细的客户化。在物理模型设计方面，将逻辑模型物理化是设计的重点，同时需要结合数据仓库平台的特点，优化数据存储，保持高效的数据装载和查询功能。

本文主要研究某银行数据仓库系统建设项目，项目基于对该银行业务及相关源系统的理解，参考 Teradata 行业模型 FS-LDM，通过针对该行具体业务特点进行了模型客户化实施，形成了数据仓库的概念模型设计，并采用面向主题的设计方法，通过对来源多样和异构的业务数据进行加工转换，并使用全行统一的逻辑语言描述具体业务，从而保证了数据的一致性和高效的数据利用性。在此基础上可满足下游多种不同的应用需求和不同的数据访问方式，真正实现了数据的高效可利用性、完整性和统一性。以该项目为依托结合本人的实际工作，本文主要研究的内容包括以下几个方面：

1. 介绍数据仓库相关理论
2. Teradata 提出的数据仓库方法论
3. Teradata 产品的技术特点
4. 结合银行数据仓库项目实施过程介绍基于 Teradata FS-LDM 的数据仓库建模方式。在逻辑模型建设方面，研究重点是模型设计时遵循的主要原则，体现出逻辑模型以模型客户化为准则的设计理念。在物理模型建设方面，研究重点是将逻辑模型物理化。根据本文所研究的某银行模型特点，在物理模型设计时采用

图形的展现方式，以面向业务的主题方法，运用统一的逻辑语言对业务进行描述，对来源多样的业务数据进行维度整合，保证数据的一致性，实现信息的集中和共享。

5. 根据实际工作经验总结银行数据仓库设计的流程规范，重点介绍模型客户化的设计理念和实施流程，包括模型客户化的产物、客户化原则、模型设计步骤等。通过对 ETL 开发工作流程的描述和代码标准化的介绍，总结出为制定系统共性度高、业务认可的标准数据所要经历的过程。

1.4 论文结构

本文主要分为六章，相关的组织结构如下所示：

第一章，绪论部分。主要介绍了论文的选题背景，现在的数据仓库在银行中研究应用的具体情况，主要介绍了本文的选题的原因以及研究意义，然后又对选题的内容以及论文的结构进行详细的介绍。

第二章，介绍相关的技术框架。主要对数据仓库的相关内容以及研究过程中所需要的 Teradata 技术进行了分析，分析了 Teradta 数据仓库的优势，阐述了 Teradata 可扩展数据仓库方法论和基本框架的内容。

第三章，结合上一章中数据仓库的基本概念和 Teradata 技术特点，分析了数据仓库模型设计方法，介绍了数据仓库模型的概念，对比较重要的名词做了相关解释，介绍了数据建模的四个层次，总结基本建模方法。以本人所参与实施的某银行数据仓库建设项目为依托，分析了该项目的主题模型和模型设计的原则和特点。

第四章，介绍了该项目的实施背景、实施范围，项目一期 EDW 的建设背景和目标，总结了模型设计的步骤、具体内容及设计依据。

第五章，据该银行的具体情况和 Teradata FS-LDM 模型，介绍模型设计体验之客户化流程的方法及工作内容，总结了工作中的遇到的问题和经验

第六章，总结和期望。主要是对本文的研究内容以及取得的成果进行了详细的总结，并对以后的工作进行了布置。

第二章 数据仓库技术分析

2.1 数据仓库发展史简介

数据仓库的萌芽阶段始于 20 世纪 70 年代，MIT 公司在一项致力于优化架构的研究中提出了将业务系统和分析系统分开的理念。针对业务系统和分析系统各自特点而采取不同的架构设计原则，将业务处理和分析处理分为不同的层次。

到了 20 世纪 80 年代中后期， DEC 公司根据 MIT 的研究理论，建立了一个新的规范，即 TA2（Technical Architecture2）规范，主要对分析系统中的四个组成部分进行了研究分析，分别是数据获取部分、数据访问部分、目录部分以及用户服务系统[2]。这个转变在整个系统框架历史上具有非常重要的意义，它将萌芽阶段的理论分析发展成为明确的系统架构并付诸于实践。同时，IBM 面临的信息孤岛问题也愈发严重，众多分立系统都有各自不同的编码方式和数据格式，使数据集成问题显得迫在眉睫。

1988 年，IBM 公司为了能够解决公司的数据集成问题，首次提出了信息仓库这个定义[2]，并于 1991 年在 DEC TA2 的基础上制定了 VITAL 规范（Virtually Integrated Technical Architecture Lifecycle）[3]，主要包括 85 种信息仓库组成部分，分别是数据抽取、变换、加载以及图形查询工具以及有效性验证等。随着信息仓库组件的确立，数据仓库的基本原则也确立了下来，它包含数据仓库的基本原理、技术架构和分析系统等内容。在 1991 年，Bill Inmon 的数据仓库理论将数据仓库提升到另外一个新的高度，他凭借 1991 出版的《Building the DataWarehouse》，被誉为数据仓库之父。书中对数据仓库的概念做了如下定义：数据仓库是一种新型的数据集合，其特点主要有四点：面向主题的（Subject Oriented）、具有集成功能（Integrated）、不可以进行更新（Non-Volatile）、能够充分反映历史变化的（Time Variant）[4]，其主要功能是支持管理决策。由此可见，数据仓库在一开始的定位就是面向企业高层管理者和业务人员的，为支持管理决策而服务的。随着应用的发展，对数据仓库的要求也上升到运营系统可以与决策支持系统共享有用的信息，更好的为企业中层管理者甚至一线操作者服务。